Abundant local interactions in the 4p16.1 region suggest functional mechanisms underlying SLC2A9 associations with human serum uric acid

Citation for published version:

Digital Object Identifier (DOI):
10.1093/hmg/ddu227

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Publisher's PDF, also known as Version of record

Published In:
Human Molecular Genetics

Publisher Rights Statement:
© The Author 2014. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/3.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.
Abundant local interactions in the 4p16.1 region suggest functional mechanisms underlying SLC2A9 associations with human serum uric acid

Wen-Hua Wei1,2,∗, Yunfei Guo3, Alida S.D. Kindt1, Tony R. Merriman4, Colin A. Semple1, Kai Wang3 and Chris S. Haley1

1MRC Human Genetics Unit, Institute of Genetics and Molecular Medicine, University of Edinburgh, Western General Hospital, Crewe Road, Edinburgh EH4 2XU, UK, 2Arthritis Research UK Centre for Genetics and Genomics, Institute of Inflammation and Repair, Faculty of Medical and Human Sciences, Manchester Academic Health Science Centre, University of Manchester, Oxford Road, Manchester M13 9PT, UK, 3Zilkha Neurogenetic Institute, University of Southern California, Los Angeles, CA, USA and 4Department of Biochemistry, University of Otago, PO Box 56, Dunedin, New Zealand

Received February 4, 2014; Revised April 16, 2014; Accepted May 6, 2014

Human serum uric acid concentration (SUA) is a complex trait. A recent meta-analysis of multiple genome-wide association studies (GWAS) identified 28 loci associated with SUA jointly explaining only 7.7% of the SUA variance, with 3.4% explained by two major loci (SLC2A9 and ABCG2). Here we examined whether gene–gene interactions had any roles in regulating SUA using two large GWAS cohorts included in the meta-analysis [the Atherosclerosis Risk in Communities study cohort (ARIC) and the Framingham Heart Study cohort (FHS)]. We found abundant genome-wide significant local interactions in ARIC in the 4p16.1 region located mostly in an intergenic area near SLC2A9 that were not driven by linkage disequilibrium and were replicated in FHS. Taking the forward selection approach, we constructed a model of five SNPs with marginal effects and three epistatic SNP pairs in ARIC—three marginal SNPs were located within SLC2A9 and the remaining SNPs were all located in the nearby intergenic area. The full model explained 1.5% more SUA variance than that explained by the lead SNP alone, but only 0.3% was contributed by the marginal and epistatic effects of the SNPs in the intergenic area. Functional analysis revealed strong evidence that the epistatically interacting SNPs in the intergenic area were unusually enriched at enhancers active in ENCODE hepatic (HepG2, \( P = 4.7\times 10^{-6} \)) and precursor red blood (K562, \( P = 5.0\times 10^{-6} \)) cells, putatively regulating transcription of WDR1 and SLC2A9. These results suggest that exploring epistatic interactions is valuable in uncovering the complex functional mechanisms underlying the 4p16.1 region.

INTRODUCTION

Human serum uric acid concentration (SUA) is the outcome of balancing production (primarily in the liver) against excretion (mostly in the kidney) (1). High SUA (i.e. hyperuricaemia) can lead to gout (2). SUA is a complex trait with estimated heritability ranging from 40 to 70% (3–5). A recent meta-analysis comprising > 140 000 individuals of European ancestry identified 18 novel loci in addition to 10 previous known that jointly explained only 7.7% of the SUA variance, of which 3.4% was explained by SLC2A9 and ABCG2 (6). The meta-analysis results reiterate the ‘missing heritability’ issue (7,8) but reinforce the impression that increasing the sample size is effective in discovering novel loci but with decreasing effects (9–11). Gene–gene interactions (epistasis)—a potential source of SUA variation, were not considered in the meta-analysis study (6). Tools for analysing

∗To whom correspondence should be addressed at: Arthritis Research UK Centre for Genetics and Genomics, Institute of Inflammation and Repair, Faculty of Medical and Human Sciences, Manchester Academic Health Science Centre, University of Manchester, Oxford Road, Manchester M13 9PT, UK. Tel: +44 1612755639; Fax: +44 161275 5043; Email: wenhua.wei@manchester.ac.uk

© The Author 2014. Published by Oxford University Press. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/3.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.
epistasis at the genome-wide level currently can only handle SNPs with precise genotypes (12–16) and thus are unable to support meta-analysis of epistasis that requires imputed SNPs with probability-attached genotypes.

In contrast to the great success in genome-wide association studies (GWAS) (attributable mostly to meta-analysis) (9), the genome-wide search for epistasis in individual GWAS populations so far has been disappointing in general (17,18). This may not be too surprising because the power of detection of pairwise epistasis is a function of the interaction effect and sample size as well as linkage disequilibrium (LD) between a genotyped SNP and underlying causal variants at both loci (rather than one locus in conventional GWAS). Overall one requires a much larger sample size (18,19) than offered in each individual GWAS population. The low power issue is amplified by the need to apply significance thresholds derived from Bonferroni correction of billions of multiple tests with consensus thresholds (like 5.0E−08 for GWAS) not yet available (20). The high-density SNP coverage of the genome that is essential to provide sufficient LD for detecting epistasis is not available in most GWAS cohorts genotyped with older, relatively low-density SNP chips (21–24), posing difficulties to both detection and replication of epistatic signals. For example, in our previous study of epistasis in SUA using small isolated populations genotyped by chips with ~300 000 SNPs, interactions involving SNPs in low LD, e.g. two SNPs located within 1 Mb on the same chromosome and with an interaction \( P \)-value (\( P_{int} \)) of \(<1.0E−05 \) (21,24). Such local interactions may exist within a gene or between neighbouring genes (30,31), and rather than capturing functional genetic interactions could potentially capture variants missing from GWAS via haplotype effects (24,32) and provide new insights into the underlying molecular mechanisms (33,34).

Both approaches require no prior biological knowledge (23) and thus can provide a useful view of interactions complementary to conventional GWAS (24).

Here we used two large cohorts included in the GWAS meta-analysis (6), i.e. the Atherosclerosis Risk in Communities study cohort (ARIC) and the Framingham Heart Study cohort (FHS) both genotyped with >500 000 SNPs, to re-examine epistasis in SUA comprehensively. We used the ARIC samples with European ancestry for discovery and the FHS cohort (excluding samples in generation one) for replication. We performed full pairwise genome scans for both cohorts using a fast tool BiForce (12) and examined SNP interactions in three categories—with and without marginal SNPs and local interactions, and using specific significance thresholds derived following the procedures previously defined (21,24).

**RESULTS**

After careful data scrutiny and quality control (see Materials and Methods), 514 662 SNPs and 9172 samples (4884 females) in ARIC and 410 947 SNPs and 5538 samples (2951 females) in FHS were used in subsequent data analyses (Supplementary Material, Table S1). SNP positions quoted in this study are based on the human genome build (UCSC hg19/NCBI 37.3). Conventional GWAS identified 166 genome-wide significant \( P <5.0E−08 \) SNP associations in ARIC (Supplementary Material, Table S2 and Fig. S1) and 75 in FHS (Supplementary Material, Table S3 and Fig. S2), allocated mostly to the SLCA249-WDR1 (4p16.1) and ABCG2 regions (4q22) in both cohorts. These results are in line with the meta-analysis (6). The lead SNP associated with SUA was rs3733588 in both cohorts (Supplementary Material, Tables S2 and S3).

Using the Bonferroni-corrected threshold of 3.8E−13 for a full pairwise genome scan in ARIC, we identified five significant epistatic SNP pairs that were well replicated in FHS when both SNPs were genotyped (as was the case for 3 of the 5 pairs, see Table 1). Each of the five pairs involved at least one marginal SNP (Supplementary Material, Table S2 and Fig. S1) and had no LD between the two SNPs. All interacting SNPs were located in an intergenic area between \( WDR1 \) and \( ZNF518B \) within the 4p16.1 region, where the top four pairs of SNPs fell into a small window of <30 kb implicating a common epistatic signal upstream of rs3733588 (Fig. 1).

Using the genome-wide threshold of 5.9E−10 for interactions involving marginal SNPs (Materials and Methods), we further identified 83 significant pairs of SNPs all mapped to the 4p16.1 region, of which 45 pairs of interactions were directly replicated (i.e. both SNPs were genotyped with \( P_{int} <0.05 \)) in FHS (Supplementary Material, Table S4). The 48 directly replicated SNP pairs (including the three in Table 1) were plotted in Figure 1, showing they were scattered mostly in the intergenic areas upstream of rs3733588.

A further assessment of local interactions (i.e. two SNPs within 1 Mb with \( P_{int} <1.0E−05 \)) found the whole 4p16.1

<table>
<thead>
<tr>
<th>chr</th>
<th>SNP1</th>
<th>pos1</th>
<th>SNP2</th>
<th>pos2</th>
<th>Dist</th>
<th>LD (( r^2 ))</th>
<th>( P_{int} )</th>
<th>( P_{int,FHS} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>rs4697924</td>
<td>10 124 239</td>
<td>rs731069</td>
<td>10 152 431</td>
<td>28.2</td>
<td>0.000</td>
<td>6.1E−14</td>
<td>NA</td>
</tr>
<tr>
<td>4</td>
<td>rs4697924</td>
<td>10 124 239</td>
<td>rs477357</td>
<td>10 152 878</td>
<td>28.6</td>
<td>0.000</td>
<td>2.2E−13</td>
<td>NA</td>
</tr>
<tr>
<td>4</td>
<td>rs4697926</td>
<td>10 124 567</td>
<td>rs731069</td>
<td>10 152 431</td>
<td>27.9</td>
<td>0.000</td>
<td>4.0E−14</td>
<td>1.5E−07</td>
</tr>
<tr>
<td>4</td>
<td>rs4697926</td>
<td>10 124 567</td>
<td>rs477357</td>
<td>10 152 878</td>
<td>28.3</td>
<td>0.000</td>
<td>1.7E−13</td>
<td>1.2E−07</td>
</tr>
<tr>
<td>4</td>
<td>rs11722989</td>
<td>10 126 139</td>
<td>rs684580</td>
<td>10 208 794</td>
<td>82.7</td>
<td>0.002</td>
<td>3.1E−13</td>
<td>6.9E−04</td>
</tr>
</tbody>
</table>

chr—chromosome of a SNP pair; SNP1, SNP2—name and position of the first (second) SNP; pos1, pos2—name and position of the first (second) SNP; Dist—distance in kb between two SNPs; LD (\( r^2 \))—linkage disequilibrium between two SNPs; \( P_{int} \)—\( P \)-value of the interaction test; \( P_{int,FHS} \)—interaction \( P \)-value of the SNP pair in FHS; NA—not directly replicated in FHS; SNPs in bold were genome-wide significant in GWAS in ARIC.
region was enriched with interaction signals (917) in ARIC (Fig. 2) as well as FHS (Supplementary Material, Fig. S3). Outside of this region, we also observed a strong local interaction between rs2622621 and rs1564481 (both SNPs within ABCG2 with $P_{\text{int}} = 6.2\times10^{-11}$, distance = 30.3 kb, $r^2 = 0.23$), which, however, was not replicated in FHS. Another SUA-associated gene tagged by local interactions in ARIC was BCAS3 on chromosome 17: rs9914370 (BCAS3)–rs758596 (TBX4) ($P_{\text{int}} = 6.0\times10^{-6}$, distance = 522.5 kb, $r^2 = 0.0$) that was not replicated in FHS either.

We then performed conditional tests of the 917 local interaction pairs observed in the 4p16.1 region in ARIC by fitting the lead associated SNP rs3433588 (additive effect only) in the background and found 27% of them with $P_{\text{int}} < 0.05$ across the region (Fig. 2). All the top five SNP pairs in Table 1 and 38 out of the 45 directly replicated significant SNP pairs (Fig. 1, Supplementary Material, Table S4) passed the conditional tests, suggesting they did not simply mirror the marginal effects of rs3433588. Similar conditional tests of the remaining marginal SNPs within the region also suggested multiple independent associations ($P < 0.05$) that were generally in low LD with the lead SNP rs3433588, except for several SNPs in relatively long range (e.g. > 100 kb) LD ($0.5 < r^2 < 0.6$) (Supplementary Material, Table S5 and Fig. S4).

Using the forward selection approach, we selected five independent marginal SNPs (i.e. rs3733588, rs874432, rs4697695, rs9291683 and rs11734783) capturing most marginal effects of the region, where the first three are within SLCA29 and the latter two are intergenic between WDR1 and ZNF518B. Conditioning on the five selected marginal SNPs, we still found ~10% (88 out of 917) of the 4p16.1 local interactions remained significant ($P_{\text{int}} < 0.05$), including only one genome-wide significant pair (i.e. rs46977708–rs16895984, conditional $P_{\text{int}} = 0.047$, $r^2 = 0.122$) (Supplementary Material, Table S6 and Fig. S5). A forward selection of the 10% significant local interactions led to three independent SNP pairs all located between WDR1 and ZNF518B (Table 2). The selected five marginal SNPs and three SNP pairs jointly explained 6.0% of the SUA residual variance—1.5% more than that explained by the lead SNP alone but only 0.3% was contributed by the two marginal SNPs and three epistatic SNP pairs in the intergenic area with the remaining 1.2% owing to the two additional SNPs in SLCA29. Nevertheless, without conditioning on the five marginal SNPs, the three epistatic SNP pairs could jointly explain 3% of the SUA residual variance.

We further imputed the 4p16.1 region (i.e. 9 900 000–10 400 000) in ARIC using the 1000 Genomes Project reference panel and tested whether the observed local interactions reflect a single untyped variant. Forty-four imputed SNPs had associations stronger than rs3733588, but the associations were not substantially different (Supplementary Material, Table S7). Using the forward selection approach described earlier, six independent imputed SNPs were selected: rs938558, rs4428284, rs4697695 (also typed), rs10489074, rs4481233 (also typed) and rs16895984 (also typed), where rs10489074 and rs16895984 are intergenic between WDR1 and ZNF518B and the rest are within SLCA29. Conditioning on the six selected imputed SNPs, 45 local interactions remained significant ($P_{\text{int}} < 0.05$) (Supplementary Material, Table S8), most also appeared in Supplementary Material, Table S6.
including the genome-wide significant rs4697708 – rs16895984. These results suggested there were local interactions independent to marginal effects.

The genome-wide significant local interactions in the 4p16.1 region (Table 1 and Supplementary Material, Table S4, 42 unique epistatic SNPs) were analysed for enrichment of active regulatory regions using an online tool HaploReg (35). We found significant enrichments of enhancer regions in ENCODE (36) HepG2 (hepatocyte, 5.2-fold enrichment, \( P = 4.7E-5 \)) and K562 (blood erythroblast, 5.2-fold enrichment, \( P = 5.0E-6 \)) cell lines (Table 3). No similar significant enrichments were seen for ENCODE enhancer annotations for a variety of other ENCODE cell types (HUVEC umbilical vein endothelial cells, HMEC mammary epithelial cells, GM12878 B-lymphocytes, NHEK epidermal keratinocytes).

The significant enrichments of enhancers in HepG2 and K562 cell lines were also observed in the analyses of GWAS marginal SNPs in the 4p16.1 region in both ARIC (Supplementary Material, Table S9) and FHS (Supplementary Material, Table S10).

Closer examination of the chromatin states predicted using the software chromHMM (37) also suggested distinct enhancer activity signals in the 4p16.1 region (the chromHMM category plot, Supplementary Material, Fig. S6), including several strong enhancers located in the two intergenic areas flanking WDR1 in the HepG2 cell line. Consistent with this, transcription factor binding-site clusters and various other features of functionally active HepG2 chromatin were also found in these intergenic areas (Supplementary Material, Fig. S6). The intergenic area between WDR1 and ZNF518B appears to be bound by transcription factors and RNA polymerase II in cell lines including HepG2 and K562 (Supplementary Material, Fig. S7). To investigate whether ChIA-PET interactions are commonly enriched in other GWAS regions, we sorted 17 680 marginal SNPs currently available from the GWAS Catalog (38) into 8817 regions each encompassing 25 kb and then counted ChIA-PET interactions within a distance of 100 kb flanking the middle point of each sorted region. We found <4% (352 out of 8817) of the sorted GWAS regions had more ChIA-PET interactions than the SLC2A9 region (i.e. chr4: 10080000–10105000, with 96 ChIA-PET interactions). This empirical analysis places the observed SLC2A9 enrichment for interactions within a small proportion of known complex trait loci identified by GWAS.

We further examined local interactions in the 4p16.1 region in ARIC female and male samples separately following the same analysis procedure (Supplementary Material, Fig. S9). Local interactions in the 30-kb window appeared relatively consistently in both genders suggesting a common enhancer activity. The most striking difference was that females had very strong (\(-\log_{10} \text{P}_{\text{int}} > 14\)) long range interactions between two WDR1 SNPs (rs4604059 and rs12498927) and two intergenic SNPs (rs7681212 and

<table>
<thead>
<tr>
<th>Table 2. Statistical construction of the genetic structure of the 4p16.1 region</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>SNP1</strong></td>
</tr>
<tr>
<td>rs3733588</td>
</tr>
<tr>
<td>rs4696695</td>
</tr>
<tr>
<td>rs1173483</td>
</tr>
<tr>
<td>rs4698000</td>
</tr>
<tr>
<td>rs6813385</td>
</tr>
</tbody>
</table>

only additive effects considered for first marginal SNPs; SNP1 (SNP2), pos1 (pos2)— name and position the first (second) SNP; dist—distance in kb between two SNPs; LD (r²)— linkage disequilibrium between two SNPs; P-value—P-value of an epistatic SNP pair; Variance explained—SUA residual variance explained accumulated; blank cells: no information required.

<table>
<thead>
<tr>
<th>Table 3. Enrichment of ENCODE enhancers by genome-wide significant local interactions in the 4p16.1 region in ARIC</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Cell type</strong></td>
</tr>
<tr>
<td><strong>ID</strong></td>
</tr>
<tr>
<td>HepG2</td>
</tr>
<tr>
<td>HUVEC</td>
</tr>
<tr>
<td>K562</td>
</tr>
<tr>
<td>HMEC</td>
</tr>
<tr>
<td>GM12878</td>
</tr>
<tr>
<td>NHEK</td>
</tr>
</tbody>
</table>

Obs.: observed; Exp.: expected; Fold: fold of enrichment calculated as Obs./Exp.; 42 unique epistatic SNPs from Table 1 and Supplementary Material, Table S4 used in the enrichment analysis.
rs11943276) near ZNF518B, which did not appear in males (Supplementary Material, Fig. S9).

**DISCUSSION**

Using ARIC and FHS, we were able to detect genome-wide significant epistasis in SUA (Table 1) based on a stringent Bonferroni-corrected threshold (20). Additional searches focusing on marginal SNP interactions with a relaxed significance threshold found more significant epistatic signals, all within the 4p16.1 region covering SLC2A9 Fig. 1 and Supplementary Material, Table S4). The significant epistatic signals identified in ARIC were well replicated in FHS, either exactly as the same SNP pair if both SNPs were genotyped (Fig. 1), or at the regional level (Fig. 2 and Supplementary Material, Fig. S3) (23-25,39).

The observations were reinforced by abundant and widespread local interactions within the 4p16.1 region in both ARIC and FHS (with ~103 000 SNPs less than ARIC) (Fig. 2 and Supplementary Material, Fig. S3). The conditional analysis results based on the lead SNP rs3773588 showed that a number of local interactions and marginal SNPs were statistically independent, supporting the hypothesis of multiple variants residing in the region (Fig. 2 and Supplementary Material, Fig. S4). This differs from the report of only one associated SNP within SLC2A9 from the meta-analysis (6), i.e. rs12498742 that is 53 kb away from rs3773588, possibly because (a) rs12498742 did not pass the quality control in this study and (b) the meta-analysis placed an additional requirement of SNP effect size reduction (~20%) to claim an independent signal in the conditional tests.

To fully assess the impact of marginal effects on local interactions in the region, we forward-selected five marginal SNPs as the additive genetic background and still found a substantial number of local interactions significant in the new conditional tests, most involving at least one epistatic SNP located in the intergenic areas flanking the WDR1 gene (Supplementary Material, Fig. S5), suggesting there might reside some regulatory elements. We showed that these local interactions did not reflect the effects of a single imputed SNP. The ‘final’ forward selection model of the five marginal SNPs and three SNP pairs further emphasized the intergenic area between WDR1 and ZNF518B, covering the two marginal SNPs and all the three epistatic SNP pairs.

Functional analyses provide strong evidence that the epistatically interacting SNPs are unusually enriched at enhancers active mainly in hepatic and precursor red blood cell types implicated in SUA (Table 3). Despite the fact that many regions of the genome may show enhancer activity in some cell type at some time, identifying cell-type-specific enhancers by integrating GWAS and epigenetic signals has become increasingly useful for functionally studying complex traits (40). The HaploReg enrichment test applied is a statistically rigorous approach for this purpose that uses a rigorously defined genomic background given all the SNPs genotyped (35). To our knowledge, the present study is the first in applying this approach to elucidating the biological basis of epistatically interacting loci and generated testable hypotheses for follow-up functional work by experimental biologists. Particularly in the WDR1-ZNF518B intergenic area, statistical interactions, enhancers, chromatin interactions between TSS of WDR1 and the enhancers collectively suggest complex mechanisms regulating SLC2A9 function, which may potentially contribute to the SLC2A9-mediated effect on gender difference in SUA levels (6). The observations that both intergenic areas flanking WDR1 are actively transcribed with many transcripts overlapping or adjacent to each other lead to the hypothesis that SLC2A9 and WDR1 may be co-transcribed or share transcription regulatory machinery. The hypothesis is intriguing as SLC2A9 gene expression may be regulated by enhancers directly targeting SLC2A9 and/or indirectly regulated by other enhancer(s) via WDR1 transcription. Further work is needed to test these hypotheses and dissect the regulatory mechanisms.

In this study, we detected no genome-wide significant epistatic signals other than those in the 4p16.1 region, reinforcing the impression that single-GWAS populations are generally underpowered for studying epistasis (24). Indeed, even in conventional GWAS, GCKR was the only locus other than SLC2A9 and ABCG2 detected significantly in ARIC (suggestively in FHS). Hence, at the level of single-GWAS populations, searching for marginal SNP interactions and local interactions under relaxed significance thresholds are perhaps more realistic approaches. For example, local interactions also captured ABCG2 and BCA33 despite not being directly replicated in FHS. In fact, a suggestive local interaction pair of rs2725227 and rs2725222 (distance = 14 kb, $P_{	ext{int}} = 1.2E-05$ in ARIC and 9.6E-03 in FHS) was near ABCG2, with both SNPs were located in PKD2, a candidate causal locus of polycystic kidney disease (41) and regulator of SUA levels (42).

Local interactions were not found for all GWAS loci, e.g. no local interactions were observed in the GCKR locus at all in this and our previous study of eight metabolic traits (24). Abundant local interactions seem more likely to be seen in regions with greater genetic heterogeneity, e.g. the human leucocyte antigen (HLA) region for auto-immune diseases (43,44), the 11q23.3 region for lipid traits (24). We also observed local interactions in other regions not harbouring associated variants across the genome (results not shown) but concentrated on the 4p16.1 region in this study. Considering the difficulty in differentiating haplotype effects from true interactions statistically (21,24), future functional work is needed to decipher any haplotypes like those in HLA (45) or real functional interactions (34,46,47).

In this and our previous studies of epistasis in quantitative traits (21,24,25,39), the trait values used for testing epistasis were the resultant residuals of a mixed model following an adapted GRAMMAR approach (48) to correct for polygenic effects and covariates including the first 10 principal components computed from the genomic relationship matrix to account for relatedness in samples. It is known that the GRAMMAR approach is conservative in conventional GWAS that consider only additive effects and the conservativeness increases as population substructure and trait heritability increase (48,49). While it is unclear whether the GRAMMAR approach remains conservative in epistatic models, the adapted GRAMMAR approach (i.e. accounting for relatedness simultaneously) seems to be not conservative in our previous study of epistasis in SUA in isolated populations (25). Nonetheless, as a precaution, one could allow additional epistatic signals out of BiForce screening to enter the full model tests and apply predefined thresholds afterwards. Such full model tests are
essential to assess the screening results as BiForce uses approximate statistical tests for interactions and treats each pair of SNPs independently (21).

In summary, pairwise genome-wide screening for epistasis in SUA allowed us to detect abundant local interactions in the 4p16.1 region that highlighted the functional complexity of the region and provided compelling insights into potential mechanisms regulating SLC2A9 functions.

MATERIALS AND METHODS

This study was approved by the institutional review board of the West of Scotland Research Ethics Service of NHS in the UK. The GWAS data of the ARIC and FHS study cohorts are provided by the NIH Database of Genotype and Phenotype via specific Data Use Certifications issued by the Data Access Committee of the National Heart, Lung and Blood Institute. Both study cohorts have been described in detail elsewhere (50–53). Only individuals with European ancestry of the two study cohorts were used in this study. Both ARIC and FHS were approved by corresponding local ethics committees and obtained written informed consent from the study participants. ARIC was genotyped with the Affymetrix 6.0 SNP chip and the FHS cohort with Affymetrix 500K and Affymetrix 50K SNP chips.

A common protocol was used to perform quality control of the genotype data in both cohorts using the GenABEL package (54) implemented in R (http://www.r-project.org/): individual call rate at 97%, SNP call rate at 95%, minor allele frequency at 2%, P-value for deviation from Hardy–Weinberg equilibrium at 1.0E–10, false discovery rate for unacceptably high individual heterozygosity at 0.01. SUA in ARIC was corrected for sex, age, body mass index (BMI), serum creatinine, hypertension treatment, and sample centre. SUA in FHS was corrected for sex, age, BMI, creatinine, hypertension treatment, renal disease status and generation (SUA in generations 2 and 3 samples measured at their second and first visit, respectively). To control relatedness, individuals that were outliers of the first three principal components computed from the identity-by-state matrix constructed using GenABEL were removed. In addition, subjects younger than 18 years old, or with BMI > 50, or with creatinine beyond the range of 3 SD of the population mean were removed from the study. After quality control, 9172 (4884 females) and 5538 (2951 females) samples, 514 662 and 410 947 autosomal SNPs were analysed in ARIC and FHS, respectively (Supplementary Material, Table S1).

Genome scans were performed for each cohort as follows: (a) the identity-by-state matrix was reconstructed and the first ten principal components were calculated and stored; (b) SUA was adjusted for covariates correspondingly and normalized using the GenABEL rntransform function and then adjusted for polygenic effects and the first ten principal components to account for relatedness using the mixed model-based polygenic function where the polygenic heritability was computed (Supplementary Material, Table S1) and the resultant environmental residuals (i.e. pgrresidual) were used as the actual trait values for association tests (48); (c) conventional GWAS analyses (i.e. assuming additive effects only) were performed using the GenABEL mmscore function (49) and the consensus threshold ($P = 5.0E−08$) (55) was used to identify marginal SNPs; (d) full pairwise genome scans using BiForce that utilizes bitwise data structures and advanced algorithms to allow high-throughput detection of epistasis (12). Genome-wide significant thresholds were derived based on the Bonferroni adjustment of actual number of tests as previously described (12,21), i.e. with 514 662 SNPs and 166 marginal SNPs identified (Supplementary Material, Table S2) in ARIC, the thresholds were $3.8E−13$ ($P = 0.05/ (514 662 × (514 662−1)/2)$) for SNP pairs identified from the full pairwise genome scan and $5.9E−10$ ($P = 0.05/((514 662−1) × 166)$) for SNP pairs involving at least one marginal SNP. We adopted the threshold of $1.0E−05$ for local interactions derived previously based on permutation (24).

Significant epistatic SNP pairs were tested for replication in FHS at the SNP level only for simplicity, i.e. claiming a replication of an epistatic pair only if both SNPs were genotyped and with $P_{int} < 0.05$ in FHS (24). Conditional tests were carried out by fitting one or multiple marginal SNPs as fixed effects in the background and then each of other SNPs or SNP pairs individually in the same way(s) as used in the genome scans and considering the SNP or SNP pair statistically independent if the conditional $P/P_{int} < 0.05$. The forward selection approach was used when multiple independent associations were available in the conditional tests: to select the most associated SNP or SNP pair (i.e. with the lowest conditional $P/P_{int}$), fit into the background and test the remaining, repeating until no more significant conditional associations were found. Variance explained was calculated using the polygenic function with marginal SNPs or SNP pairs fitted as fixed effects.

We imputed the 4p16.1 region (from 9900 to 10400 kb) based on 9172 samples and 260 typed SNPs in ARIC using IMPUTE2 (56) and the 1000 Genomes Project reference panel (phase1 integrated variant set v3). We used SNPTEST (v2.5) (57) to test associations of 2610 imputed SNPs (minor allele frequency > 0.01) with the same SUA trait in the frequentist additive model using genotype dosages. We used PLINK2 (https://www.cog-genomics.org/plink2/) to take the best genotypes of the imputed SNPs and then performed forward selection and conditional tests in R as described earlier.

GWAS marginal SNPs and genome-wide significant epistatic SNPs within the 4p16.1 region were analysed for enrichment of ENCODE (36) cell-type-specific enhancers using the online tool HaploReg (http://compbio.mit.edu/HaploReg) that tests enrichment based on a rigorously defined genomic background (i.e. all the SNPs genotyped) (35), with LD information ($r^2 > 0.8$) from the 1000 Genomes Project and a background set of Affymetrix 6.0 SNPs. ANNOVAR (58) and UCSC genome browser (59) were used for functional annotation of SNPs within the region to identify regulatory signals associated with these loci. Enlight (http://enlight.usc.edu) was used to visually inspect the relationship between LD and regulatory signals.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at HMG online.

ACKNOWLEDGEMENTS

We thank the editor and anonymous reviewers for their valuable comments. We are grateful to Dr John Bowes for providing
assistance in the imputation work. We acknowledge data access to the Atherosclerosis Risk in Communities study (ARIC, phs000901.v1.p1) and the Framingham Heart Study (FHS, phs000007.v18.p7) granted by the National Heart, Lung, and Blood Institute (NHBLI) Data Access Committee via Database of Genotypes and Phenotypes (http://www.ncbi.nlm.nih.gov/gap). The Framingham Heart Study is conducted and supported by NHBLI in collaboration with Boston University (Contract No. N01-HC-25195). This manuscript was not prepared in collaboration with investigators of the Framingham Heart Study and does not necessarily reflect the opinions or views of the Framingham Heart Study, Boston University, or NHBLI.

The Atherosclerosis Risk in Communities Study is carried out as a collaborative study supported by NHBLI contracts N01-HC-55015, N01-HC-55016, N01-HC-55018, N01-HC-55019, N01-HC-55020, N01-HC-55021, N01-HC-55022, R01HL087641, R01HL59367 and R01HL086694; National Human Genome Research Institute contract U10HG004402; and National Institutes of Health contract HHSN268200625226C. The authors thank the staff and participants of the ARIC study for their important contributions. Infrastructure was partly supported by Grant Number UL1RR025005, a component of the National Institutes of Health and NHLBI Roadmap for Medical Research.

Conflicts of Interest statement. None declared.

FUNDING

W.H.W., C.A.S. and C.S.H. are supported by the UK Medical Research Council University Unit Strategic Partnership Funding to MRC Human Genetics Unit, University of Edinburgh. W.H.W. is partially funded by Higher Education Funding Council for England (HEFCE). Y.G. and K.W. are supported by NIH Grant R01 HG006465. BBSRC Travel Grant (BB/K004964/1) initiated the collaboration with K.W. Funding for open access charge: the UK Medical Research Council Core Fund. Funding to pay the Open Access publication charges for this article was provided by the UK Medical Research Council University Unit Strategic Partnership Funding to MRC Human Genetics Unit, University of Edinburgh.

REFERENCES